CHALLENGES IN LONG-TERM DATA STEWARDSHIP

Ruth Duerr, Mark A. Parsons, Melinda Marquis, Rudy Dichtl, Teresa Mullins

National Snow and Ice Data Center (NSIDC)
449 UCB, University of Colorado
Boulder, CO 80309-0449
(rduerr, parsonsm, marquism, dichtl, tmullins)@nsidc.org
Telephone: +1 303-(735-0136, 492-2359, 492-2850, 492-5532, 492-4004)
Fax: +1 303-492-2468

1 Introduction

The longevity of many data formats is uncertain at best, and more often is disturbingly brief. Maintenance of backwards compatibility of proprietary formats is frustratingly limited. The physical media that store digital data are ephemeral. Even if the data are properly preserved, the information that allows the data to be searched and which maintains the context of the data is often lost, threatening data utility. These are only a few of the formidable problems that threaten the long-term preservation and long-term use of digital data.

Over the past decade, much has been written about the problems of long-term digital preservation (see for example [14], [15], [32], [38], and [39]). Many approaches or strategies to address these problems have been proposed (see for example [7], [10], and [32]), and a number of prototypes and test beds have been implemented (see for example [44]). No one has developed a comprehensive solution to these problems. In fact, there may not be a single solution.

Most of the literature applies directly to the needs of libraries, museums, and records management organizations. Only rarely are issues related to preservation of science data discussed directly. Stewards of scientific data often face much different issues than the typical library, museum, or records archive. Some issues are simpler others more complex.

In this paper, we provide a brief history of data stewardship, particularly science data stewardship, define long-term stewardship; and discuss some of the problems faced by data managers. We describe a broad array of data stewardship issues, but we will focus on those that are particularly amenable to technological solutions or that are exacerbated when archives are geographically distributed.

2 A Brief History of Scientific Data Stewardship

A cursory review of scientific data stewardship as a discipline distinct from document preservation or records management suggests that it is a fairly recent concept. For most of human history, what little scientific data existed was recorded in notebooks, logs or maps. With luck, a library or archive would collect and preserve these logs and maps. The archives may have been maintained by the church, a professional society, or perhaps were established through government regulation, but it was generally an ad hoc affair. Unless a potential data user was already aware of the existence and location of certain "data set," it was extremely difficult to find and access the data.

The establishment and growth of academic and public libraries in more recent centuries greatly improved data preservation and access. Libraries were at the forefront of new data cataloging, indexing, and access schemes; librarians were critical data stewards. Yet the "data" were still primarily in the form of monographs and logbooks, and, logically, libraries focused more on books, journals, and other publications more concerned with data analysis. (Maps may have been as readily archived as books and journals). It wasn't until the establishment of the World Data Centers (WDCs) in 1957-1958 that the concept of a publicly funded facility specifically charged with providing data access and preservation became prominent [1].

The World Data Center system originally archived and distributed the data collected during the International Geophysical Year [1]. The data in question were generally small in volume and certainly not digital, but the concept that an institution would focus on the preservation and distribution of raw data as opposed to the interpretation of those data was revolutionary. Furthermore, the WDCs were organized by disciplines such as glaciology and meteorology. This helped reinforce an association between discipline-specific science and data stewardship.

Since then, the number of discipline-specific data centers has grown. In the US a total of nine national data centers were established, primarily sponsored by NOAA, DOE, USGS and NASA [2] to archive and distribute data in disciplines such as space science, seismology, and socioeconomics. The development of these world and national centers made finding relevant data a little simpler. Now there was likely to be an organization that could be queried if only by mail or telephone. If they couldn't provide the data directly, they were usually able to provide references to other places to look.

Local and state governments, universities, and even commercial entities have continued the trend and established a variety of data centers, typically organized around disciplines or subject areas as diverse as "advertising" [3] or "cancer in Texas" [4]. The Federal government again made a significant contribution in the early 1990s when NASA established eight discipline-specific Distributed Active Archive Centers (DAACs) to collaboratively archive and distribute data from NASA's Earth Science Enterprise (ESE).

In some ways the DAAC system followed the model of the distributed and discipline-specific World and National Data Centers, and NASA typically collocated the DAACs with already established data centers [2]. However there are some key differences in the approach. On one hand, DAACs are intended to only archive and distribute data during the most active part of the data life cycle. The DAACs are to transfer their data to a permanent archive several years after each spacecraft mission in the ESE program ends, but the details of this transfer are yet to be finalized. On the other hand, an early and important goal of the ESE was to make finding and obtaining Earth science data simpler than it had been.

The DAACs are part of a larger system of remote sensing instruments and data systems called the Earth Observing System (EOS). They are linked together through the EOS Data and Information System (EOSDIS) Core System (ECS), which provides tools and hardware to handle ingest, archival, and distribution of the large volumes of data generated by EOS sensors and heritage data sources. An important component of ECS is an electronic interface that allows users to search and access the holdings of all of the DAACs simultaneously. This interface was initially developed as an independent client that users would install on their own machine, but shortly after ECS development started, the first web browsers became available. This led to the development of the EOS Data Gateway (EDG), a web-based search and order tool. Currently the EDG allows search and access to DAAC data as well as data located at several data centers scattered around the world.

What is important to note about ECS and the DAACs is that it was arguably the functional beginning of new model of data management where data archival was geographically distributed, but search and order were centralized. It is also notable that this was a newly comprehensive effort to acquire, archive, and provide access to a very large volume of data but there is still no concrete plan for the long-term disposition of the data. Both these trends—centralized access to decentralized data and inadequate planning for long term archival—continue today. Indeed NASA is moving further away from a data center approach with its new Strategic Evolution of Earth Science Enterprise Data Systems (SEEDS) [12].

Of course, the World Wide Web has been a major driver in the increased decentralization of data storage. Furthermore, improved search engines theoretically make it easier than ever to find data. We have even heard it suggested that Google may be the only search engine needed. General search engines, however, provide little information to help a user determine the actual applicability or utility of the data found. Little of the information currently available on the web has been subject to the levels of peer-review, copyediting, or quality control traditionally done by data managers or library collection specialists [18]. Finally, no mechanism ensures the preservation of much of the information available via the Web. Often web sites cited in a paper are no longer active mere months after the publication of the paper [43].

There are many efforts underway to address some of the issues inherent in distributed Earth-science data systems including the overall Web. Some examples of centralized search tools for distributed scientific data include:

- NASA’s Global Change Master Directory (GCMD) (http://gcmd.nasa.gov)
- The Distributed Oceanographic Data System (http://www.unidata.ucar.edu/packages/dods/index.html)

- The Alexandria Digital Library Project (http://alexandria.ucsb.edu/)

- The National Spatial Data Infrastructure (NSDI) (http://www.fgdc.gov/nsdi/nsdi.html)

Some of these efforts predate the World Wide Web, and some like the GCMD are strictly search tools, while others such as the NSDI attempt (with mixed success) to provide actual data access.

As data managers at the National Snow and Ice Data Center (NSIDC), we are primarily concerned with Earth science data, but we should note that many of the issues we will discuss apply to a variety of disciplines. Based on some of our experience at a session on "Virtual Observatories" at the Fall 2003 meeting of the American Geophysical Union, it seems that non-Earth Science related disciplines sometimes lag behind the Earth sciences in the management of their data. Mechanisms for simultaneously searching and accessing data stored at multiple distributed data centers may not exist. For example, no equivalent to the GCMD or EDG currently exists for the solar or space physics community. This situation is rapidly changing. Numerous groups are working on virtual observatory concepts, which in some ways are reminiscent of the EOS DAAC system described earlier.

We should also be aware of the growth of private records management companies. It is certainly possible for commercial entities to address some of the issues of modern data stewardship, but very little research has been done to accurately quantify the necessary costs of a distributed data management infrastructure. Nor have there been any significant efforts to do a cost-benefit analysis of the various components of such a structure [17]. This is especially true in the international context, where not only is distributed data management more challenging, but cost models become more difficult. For example, different countries have data access and pricing policies that are rooted less in economics than in political or philosophical issues such as the right for citizens to access government documents (See [17] and [16] for examples.).

In the following sections, the challenges of providing distributed data discovery and access, while adequately addressing long-term stewardship will be discussed. NSIDC's more than 25-year history as:

- A World Data Center

- Part of a NOAA cooperative institute

- A NASA Distributed Active Archive Center (DAAC),

- NSF's Arctic System Science (ARCSS) Data Coordination Center ADCC) and Antarctic Glaciological Data Center (AGDC)

- A central node for the International Permafrost Association's (IPA) Global Geocryological Data System (GGD)

will serve as one source of examples.

3 Long-Term Stewardship Defined

Within the data management field, “long-term” is typically defined as:

> *A period of time long enough for there to be concern about the impacts of changing technologies, including support for new media and data formats, and of a changing user community, on the information being held in a repository[5].*

Given the current rate of technological change, any data-generating project or program with a duration of five or more years should be considered as long-term and will need to take changes in technology into account.

Stewardship, especially data or scientific stewardship is more difficult to define. Of the 107 results recently found with a Google search of the phrase “scientific stewardship,” very few (primarily NOAA sites, a few religious sites, and one lumber company) actually defined what the phrase meant in their context. These concepts are relatively new and do not show up in standard information science dictionaries or encyclopedias.

The term data stewardship was used in the early 1990s by the Department of Defense in DOD Directive 8320.1-M.1, which defined data administration as “the person or group that manages the development, approval, and use of data within a specified functional area, ensuring that it can be used to satisfy data requirements throughout the organization” [40].

Two other relevant definitions can be found in the literature. The first comes from the vision statement from a workshop sponsored by NASA and NOAA which states that long-term archiving needs to be a “continuing program for preservation and responsive supply of reliable and comprehensive data, products, and information … for use in building new knowledge to guide public policy and business decisions” [11]. The second definition was presented by John J. Jensen of NOAA/NESDIS at the 2003 IEEE/NASA Mass Storage Conference, as “maintaining the science integrity and long term utility of scientific records” [45].

Both definitions associate scientific stewardship with data preservation as well as access or use in the future. These dual needs are also recognized in the library and records communities (see for example [44] and [46]). Beyond simple access to the original science data, good science stewardship has been shown to allow future development of new or improved products and for use of data in ways that were not originally anticipated [11]. To support these uses however, extensive documentation is needed including complete documentation about the characteristics of the instrument/sensor, its calibration and how that was validated, the algorithms and any ancillary data used to produce the product, etc. [11] and [12]. This level of associated documentation goes well beyond the typical metadata needs of library or records materials.

4 Data and Metadata Related Challenges

4.1 Open and Proprietary Data and Metadata Formats

The challenges of preserving information for the long term when it is stored in a proprietary format (e.g., MS Word) have been described elsewhere [6]. Commercial pressures do not allow companies to maintain backwards compatibility with each new release for very long. This leaves a very narrow window of opportunity for the information to be migrated to a newer version of the format or a different format, with the attendant risk of loss of functionality or information with each migration.

In the science stewardship realm this may not seem like a large concern since data are still often stored as ASCII tables, flat binary files or one of an increasing number of community standard formats (e.g., shapefiles, HDF-EOS 4). However, much of the associated information about the data - the information that will be needed decades later to allow reanalysis or reprocessing or to allow the development of new products - may very well be stored in a wide variety of proprietary formats (e.g., CAD files, MS Word document).

Even when the data are stored in a non-proprietary format (e.g., CDF, net-CDF or HDF-EOS), the data cannot be maintained forever in their original format. Even so-called standard formats evolve with changes in technology. For example, much of the data stored in the typically petabyte-scale archives of the NASA DAACs, are stored in either HDF-EOS 2.x or HDF-EOS 5.x formats (there are no 3.x or 4.x versions). HDF-EOS 5.x was developed as technological changes mandated entirely new data system architectures incompatible with HDF-EOS 2.x. While tools are available to help users migrate data from the 2.x version to the 5.x version, the new version is not entirely backwards compatible. NASA is currently committed to funding maintenance of both versions [8], but it is not clear whether maintenance will continue once the data are transferred to another agency for long-term archival.

Format evolution can cause particular problems in the Earth sciences where it is necessary to study long data time series in order to detect subtle changes. For example, NSIDC holds brightness temperature and derived sea ice data from a series of passive microwave remote sensing sensors. This is one of the longest continuous satellite remote sensing time series available, dating back prior to 1978. NASA is continuing this time series with a new higher resolution sensor, the Advanced Scanning Microwave Radiometer (AMSR), aboard the Aqua spacecraft. This is an exciting new addition, but scientists and data managers must work to tie the AMSR data into the existing time series. Not only will there be the normal, expected issues of intercallibrating different but related sensors, but someone will likely need to do some data format conversion. The currently intercallibrated historical data is available in flat binary arrays with ASCII headers while the AMSR data is available in HDF-EOS.

Issues such as these have resulted in a call by some for the establishment of a digital format archive [9], while others have called for conversion to a “Universal Data Format” or other technology-independent-representation upon archival (see for example [10], [13], and [32]). Both of these options require additional research according to a recent NSF-DELOS report [14]. They also increase the need for good metadata describing data format transformations and how these transformations may affect the utility of the data.

4.2 Which Standards and What Metadata?

One of the lessons learned from the ESE experience is that “community-based standards, or profiles of standards, are more closely followed than standards imposed by outside forces” [12]. Developers of the ECS system recognized that having all of the data from the entire suite of satellites and sensors in the same format would simplify user access. After consideration of several potential formats, NASA settled on the HDF-EOS, a derivative of the HDF format standard [8]. A variety of user and producer communities rebelled. As a result, while much of the data stored in the ECS system is stored in HDF-EOS format, there are a number of products, notably the GLAS data stored at NSIDC, that are not in HDF-EOS format.

In addition to the recognition that user community involvement is necessary for successful standards development and adoption, the other important concept from the quote above is the notion of a standards profile, “a specific convention of use of a standard for a specific user community” [12]. It is typically not enough to say that a particular format standard is being used (e.g., HDF or netCDF); it may be necessary to define specific usage conventions possibly even content standards acceptable to a given user community in order to ensure interoperability. These specific conventions or profiles may vary from community to community.

Probably one of the most overworked expressions in the IT industry is “Which standards? There are so many to choose from.” It is ironic that not only are there so many standards of a type to choose from; but also that there are so many types of standards about which one must make choices. In the data stewardship realm it is not enough to think about data preservation and data access format standards; one must also think about standards for metadata format and content, documentation format and content, interoperability, etc. For metadata, the question is compounded further by the need to distinguish the type of metadata under discussion, e.g., metadata for data discovery, data preservation, data access, etc.

The Open Archival Information System (OAIS) Reference Model [5] provides an information model (see Figure 1) that describes the different kinds of information needed in order to ingest, preserve and provide access to digital or analog objects. The model appears to be gaining some acceptance in the library and archive communities. It is based on the concept of an Information Package that can be found by examining its associated Descriptive Information. The components of the Information package itself are:

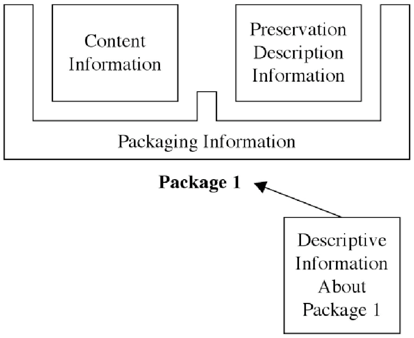
Content Information
Preservation Description Information
Packaging Information
Package 1
Descriptive Information About Package 1

Figure 1: Information Package Components and Relationships [19]

- Content Information-containing both the data object to be preserved as well as enough Representation Information for a targeted user community to understand the data object's structure and content. For science data this would include identification of the data format and any associated profile.

  Of the two components, structure and content, content is more difficult to obtain. The science community has become so specialized that community-specific jargon and underlying assumptions are pervasive. Capturing and documenting these, so that others outside that very small peer group can understand and use the data, is challenging.

  In a very distributed environment, such as the virtual observatories of the future, there will be many thousands or millions of Data Objects preserved in many different places, all of which have the same data format or even the same standard profile. It would be impractical to store the same format information with each object. This may bolster the argument for establishing centralized data format repositories [9], but would require considerable coordination to be successful.

- Preservation Description Information (PDI)-containing the information needed to preserve the object for the long-term. The PDI is comprised of 4 components:

  - Provenance, or the history of the Data Object. In the science arena, this involves tracking the processing history, what input products were used, what version of which algorithms were used, what ancillary information was used for calibration and validation, as well as a host of other instrument-related information. It also includes information about when and where the data were created, processed, and acquired; as well as who was responsible for their creation and what changes have taken place since.

- Reference Information-including information needed to identify this object from the universe of other objects. Much has been written about the need for persistent and unambiguous identifiers (see for example [32], [34], and [15]) and various communities have proposed standards for these (see for example [33] and [13]). A key finding is that in a distributed environment a name issuing authority is needed to prevent naming collisions [32]. In the science community, a hierarchy of identifiers is typically needed. For example, in the ECS system, Earth Science Data Types (ESDT's) are used to identify a data set, or class of objects, while granule ids are used to identify specific objects within the set.

- Fixity Information-documenting the methods used to ensure that the object hasn't been changed in an undocumented manner. This typically includes Cyclic Redundancy Check (CRC) values, digital signature keys, etc. This topic is addressed separately later in this paper.

- Context Information-documenting why this object was created and how it relates to other objects.

While the OAIS reference model discusses the types of metadata or information that must be gathered in order to preserve an object, it leaves the actual definition of that metadata to the individual archive or archive group. Several groups within the library community have independently developed their own preservation metadata specifications (see [21], [22], [23], and [24]) and have recently come together under the joint auspices of the Research Libraries Group (RLG) and the Online Computer Library Center (OCLC) to develop an OAIS-based metadata framework that "could be readily applied to a broad range of digital preservation activities" [25].

While providing a useful starting point for the science community, the OCLC/RLG framework is not adequate for preserving science data. The science community is typically more interested in preserving information about how the data were created than in preserving any particular presentation mechanism. This is to be expected given the different kinds of uses to which patrons of libraries and science users put the materials accessed. Typically a library patron expects to experience the material using his or her senses; to read, listen to, touch, or watch; but not to transform the materials accessed. Scientists typically access data so that it can be transformed, analyzed, used as an input to a model or new product, compared with other data, etc. Changes in presentation format over time as technology, programming languages, and visualization tools change, are not that important – the important things are the original bits and their meaning.

In the earth science realm probably the most relevant metadata standard is the "Content Standard for Digital Geospatial Metadata" established by the Federal Geographic Data Committee (FGDC) [19]. President Clinton mandated that federally funded geospatial data (i.e., most Earth science data including ESE data) adhere to the FGDC standard in an 11 April 1994 executive order [20]. The FGDC standard "was developed from the perspective of defining the information required by a prospective user to determine the availability of a set of geospatial data; to determine the fitness and the set of geospatial data for an intended use; to determine the means of accessing the set of geospatial data; and to successfully transfer the set of geospatial data" [19]. As such, there is some but not complete overlap with the kinds of metadata called for by the OAIS reference model. Much of the preservation metadata called for by the OAIS model is not part of the FGDC standard.

In the international standards realm, the equivalent to the FGDC standard is the ISO 19115 standard [26]. Like the FGDC standard, the ISO standard is meant to facilitate, discovery, assessment for use, access and use and, like the FGDC standard, does not address much of the preservation metadata of the OAIS reference model. The FGDC has developed a draft "cross-walk" between the FGDC and ISO 19115 standards which will help FGDC-compliant users also become ISO 19115 compliant users.

Both the FGDC and ISO 19115 standards are content standards, not implementation standards, yet organizations must choose implementation options. Consensus seems to be building that Extensible Markup Language (XML) should be the implementation standard for metadata. ISO Technical Committee 211 is in the process of developing an UML implementation standard for the ISO 19115 metadata standard that will include an associated XML schema. XML is also the recommendation of the National Research Council [2].

4.3 Preservation vs. Access

Users want data in formats that are easy to use. The desired format may be a community-based standard, or it may be a format that is compatible with other related data sets. Furthermore, our experience at NSIDC shows that users usually need spatial or temporal subsets of large collections and may need to combine several products. They may also need the data in a different grid or projection than the original data. In other words, the utility or essence of science data is not strongly associated with any particular access format. Indeed, many different formats, grids, and projections may need to be supported at any given time. This is significantly different from other disciplines concerned with digital preservation where it is often essential to preserve the essence of the original experience for multimedia digital objects such as movies, applications, or interactive web sites. In the Earth science community it makes more sense to consider preservation and access formats independently. Access formats are likely to change quickly over time, while preservation formats should be more stable.

There are similar issues with preservation and access metadata. There are advantages to preserving the metadata with the actual data (see section 4.4), but much of the metadata is only relevant to the archivist. The preservation-specific metadata should probably be separated from the data upon delivery to the user to minimize user confusion.

Some have called for completely separate storage of preservation and access data [13]. However, storage of multiple copies of the data is unaffordable with large data sets or when there are multiple access formats. In many cases, the issue may be how to afford preservation of even a single copy of the data! There is some agreement that the best storage format is a simple stream of bytes with adequate "representation information" (to use OAIS vernacular) to extract the necessary data object (see [7], [13], and [32]). The challenge then becomes how to adapt to changing desires for new access formats and essentially to "format on the fly."

Promising new developments in database technology that allow databases to better handle spatial information and larger objects may help address some of the data access issues, but they may actually exacerbate the data preservation problem. Not only is it impractical to store a byte stream in a database; you must also archive the representational information stored in the database. You have to worry about the evolution of database applications and schemata in addition to the other issues we are illustrating. Similarly, improved network access technologies show great promise, but they do not address the preservation problem.

4.4 Separation Issues

As discussed previously, substantial metadata or documentation is needed to allow proper understanding and reuse of data in the future. Often a simple data format is used where the documentation is kept separately (e.g., ASCII tables or flat binary files with associated readme files). This results in the risk that the metadata may be separated from the data as the data spreads through the user community. In the simplest case this might result in a user having problems understanding or reading the data. In the worst case, it might result in unintentional use of the data in non-scientifically supportable ways or the unintentional or deliberate modification of the data. Data and metadata separation can even happen if simple techniques such as the UNIX "tar" command are used to force the original user to download both the data and metadata. What happens to the individual files in the package after they have been untarred is out of the data center's control.

Some data formats such as HDF-EOS embed some of the metadata within the actual data files; but this presents a new array of problems, as discussed above. The data format may require special tools that change over time and may be incompatible with other formats. The reduced flexibility inherent in these encapsulation schemes can raise data preservation costs and may even be considered poor customer service.

In today's truly distributed environment, where the data and metadata often start out geographically separated, the issue is exacerbated. For example, NSIDC is often tasked with creating metadata and advertising data products held by external groups, typically individual investigators or investigator groups within a larger program. NSIDC provides data access as if the data were managed and stored locally. When a user requests these data, they are pointed to the provider's site.

Simply maintaining the links for these "brokered" products is a substantial challenge. For example, in 1998, NSIDC in collaboration with the International Permafrost Association (IPA) released a data compilation CD called the Circumpolar Active-Layer Permafrost System (CAPS) [41]. This was a major milestone for the permafrost community, a central element of the IPA's newly established Global Geocryological Data (GGD) system. The

CD contained over 50 data sets and nearly 100 references to other data sets held by different “nodes” of the GGD system. Unfortunately, funding for the GGD did not continue past 1998. It wasn’t until 2002, when a new initiative started to create an updated version of the CD (now a three CD set [42]), that any maintenance of the data from the 1998 version resumed. Regrettably, dozens of the distributed or “brokered” products were no longer readily available. NSIDC has plans to try and track down or “rescue” some of these data sets, but the four- to five-year time lag and the globally distributed nature of the data sets will make it very challenging. This illustrates the need for nearly constant tracking of distributed data to ensure its continued availability, or a clear and usable means (with incentives) for providers to provide updates to any central access point.

4.5 Data Security and Integrity

Ultimately, keeping track of data and metadata becomes an issue of data integrity. Scientists need to trust the validity of the data they use. They need to know that the data came from a scientifically reputable source and that the data have not been corrupted in any way.

Scientific integrity is an ill-defined concept, but it is rooted in the scientific method. Experiments must be repeatable. Results from experiments should be published in peer-reviewed literature. The data and information used in the experiment must be specifically acknowledged and generally accessible when possible. Traditionally this is handled in the literature through a system of formal citations. But while methods for citing information sources are well established and traceable, methods for citing data sources are more variable. Historically, with small non-digital data sets, the data may have been published directly in a journal or monograph that could specifically be cited. This was not an entirely consistent process, though, and as data sets have grown, authors have adopted different methods for acknowledging their data sources. Some authors may provide a simple acknowledgement of the data source in the body of an article or the acknowledgements section. Other authors may cite an article published by the data provider that describes the data set and its collection.

As publishers of data, we at NSIDC have found these historical approaches lacking. General data acknowledgements are difficult to trace, are often imprecise, and sometimes do not acknowledge the true data source. For example, an acknowledgement of “NSIDC’s SSM/I sea ice data” could actually refer to one of several different data sets and it makes no reference to the actual scientists who developed the sea ice algorithm. Citing a paper about the data is better, but in many cases such papers may not exist, they may only describe a portion of the data set, or their description may not be relevant to the new application of the data. In any case, it is not clear how to actually acquire the data—a necessary step if an experiment is to be repeated. We recommend that users cite the actual data set itself, much as they would a book or journal article. The “author” is typically the data provider or the person who invested intellectual effort into the creation of the data set (e.g., by creating an algorithm), while NSIDC or other archive that distributed the data might be considered the publisher. It is also crucial to include publication dates to distinguish between different versions of related data sets. In any case, we try and provide a specific recommended citation for every data set we distribute. Although we have met some sporadic resistance from occasional providers who wish only for their papers to be cited, this approach has become broadly accepted. It is the approach specifically recommended by the International Permafrost Association ([41], [42]), and has generally been accepted by the other NASA DAACs.

This formal citation approach works well when there is a clear and reputable data publisher even in a distributed environment. But the distributed environment may provide additional challenges, especially if data sources are somewhat ephemeral or hard to identify. For example, in a peer-to-peer system, the access mechanism needs to specifically identify the different peers and possibly some assessment of their stability. This is somewhat different than peer-to-peer systems in other areas such as music where users generally don't care where the music came from as long as it is the piece they wanted. With the rise of electronic journals we have also heard informal discussion of including the actual data used in the publication itself. Although this approach obviously includes many of the same data preservation challenges already discussed, it is an intriguing concept worthy of further exploration.

Once the scientific integrity of a data set has been assured, assurance is needed that the data received is what was expected. Several authors discuss the use of public/private key cryptography and digital signatures as methods for ensuring the authenticity of the data (see for example [35] and [36]). Lynch points out that we know very little about how these technologies behave over very long times and that, as a consequence, information about evolution of these technologies will likely be important to preserve [37].

For a scientist to be able to trust that the data have not been changed the scientist must be able to trust that the preservation practices of the source of the data are adequate: that archive media are routinely verified and refreshed, that the facilities are secure, that processes to verify and ensure the fixity of the data are operational, that geographically distributed copies of the data are maintained as a protection against catastrophe, and that disaster recovery plans and procedures are in place. To verify these practices, the RLG/OCLC Working Group on Digital Archive Attributes suggests that a process for certification of digital repositories be put in place [34]; while Ashley suggests that administrative access to data and metadata be "subject to strong proofs of identity" [36]. Once again, a distributed data environment may make implementing these suggestions more difficult.

4.6 Long-Term Preservation and Technology Refresh

A continual theme in this paper is how the speed of technological change presents a major challenge for preserving data over the long term. As a recent report by the National Science Foundation and the Library of Congress puts it "digital objects require constant and perpetual maintenance, and they depend on elaborate systems of hardware, software, data and information models, and standards that are upgraded or replaced every few years" [15].

Given the size and immediacy of the problem, it is not surprising that many different solutions have been proposed and prototyped. The literature is extensive and well summarized elsewhere (see [38] and [39]). At this point, the three most commonly used methods are normalization, migration, and emulation [15]. Normalization involves selection of a few "technology independent" standard formats, and conversion to these formats on ingest. Migration involves transferring data to new technologies before the existing technology becomes obsolete. Emulation involves recreating the original operational environment on current technologies.

To date, NSIDC has successfully employed data migration several times as media technologies have changed, and has used simple interface emulation to protect processing systems from changes in archive technology. With the growing need to preserve substantial ancillary documentation and the development of content databases it is likely that a combination approach involving normalization of incoming ancillary materials, continued migration of both the data and metadata, as well as emulation or other strategies will be needed in the future.

4.7 Size Does Count!

Finally, a word about small data sets. Much of the focus of this paper and the scientific community in general is on how to deal with very large data sets. It is easy to overlook the fact that the metadata needs for science data are more or less independent of data set size. For example, NSIDC currently publishes more than 400 data sets. Of these only 2% occupy a significant fraction of our archive. The remaining data sets are very small, typically KB or MB sized; not GB or TB. Yet the majority of metadata development resources are needed for these small data sets. Automated metadata generation tools could certainly help mitigate this resource demand, but it seems unlikely that the tools could fully address the diversity of scientific data sets currently available.

5 Scientific Stewardship Related Challenges

5.1 Maintaining Science Understanding Over Time

As discussed earlier, scientific data stewardship involves maintaining the scientific integrity of the data in order to facilitate the development of new knowledge. It is not sufficient for the data simply to be available. One must also work to facilitate future research by making data access and support simple and responsive. One must also take steps to ensure that the data is not misapplied. This implies that scientists, i.e., the data users, need to be directly involved in data stewardship. In some ways, this may be easier in a distributed environment if the scientific data provider is directly responsible for ensuring the preservation of the data, such as in the SEEDS model. On the other hand, data providers sometimes make inappropriate assumptions about the knowledge of the data user community. A good data manager, equipped with the right tools, should be working closely with the data provider to uncover any known limitations in the data.

For example, it may be self evident to developers of sea ice detection algorithms for passive microwave remote sensing that their methodology is well-suited to detection of trends in ice concentration over time but ill-suited for representing precise local ice conditions on a given day. This may not be apparent to a biologist who uses a near-real time passive microwave derived product to associate sea ice conditions in the Beaufort Sea with polar bear migration. While this is an extreme example, it highlights the need for scientists and data managers to work closely together to carefully document and track new and potentially unexpected uses of the data. It is also important to realize that the risks of inappropriate data applications could increase over time.

Of course data can also be improved. New algorithms, new calibration methods, and new instruments may be developed. In Earth science in particular, it is important to detect variability over long time periods. This means that different instruments must be intercallibrated to ensure a consistent time series, i.e. we need to be able to ensure that any changes we detect in a data stream result from actual geophysical processes not just changes in instruments or algorithms. This again requires collaboration between the data manager and the scientist. This is certainly possible in distributed environments, but mechanisms should be established to ensure that information about data harmonization and improvements are readily available to users. Traditionally, this was the role of the data steward or data manager (see, for example, [29]). It is less clear how this would work in a distributed environment, but knowledge bases and data mining systems are likely to contribute.

On a related note, to ensure maximum scientific understanding of an issue, data and support services need to be readily available to as many users as possible [11]. This is necessary to ensure all possible scientific ideas are explored and that scientific experiments can be duplicated. The necessary broad access may be better realized in a distributed data model, but only if the challenges in section four are addressed. Again this will require close interaction with the users.

Historically, NSIDC has addressed these scientific issues by working closely with its data providers and by having scientific data users and developers on staff. This becomes a less practical approach in a distributed data environment where data may be held and distributed by individuals and institutions with varying levels of scientific and data management expertise. It will become increasingly important to formally address the relationship of data managers and scientists as new distributed data management models are developed.

5.2 Decisions, Decisions, Decisions - Deciding What Data to Acquire and Retain

One of the most difficult decisions in data archival is which data to acquire and keep and which data to throw away. Although, there is still no effective business model that demonstrates the costs and benefits of long-term data archival [15], it is clearly impractical to keep all data for all time. That said, we need to recognize that many data sets often have unexpected future applications (see [11] for examples). A simple approach would be to archive a very low level of the data along with the necessary algorithms to process the higher level products. However, this must be viewed only as a minimum since it does not allow for the necessary simple and broad access described above.

It is probably not possible to describe any one infallible data acquisition and deposition scheme. However, any data stewardship model must explicitly include a method for development of such a scheme for different types of data and user communities. These schemes must explicitly include knowledgeable and experienced users of the data who are directly involved in generating new products and data quality control [11].

5.3 Upfront Planning

Our experience at NSIDC has shown that by working with the scientists and data providers early in an experiment or mission, ideally before any data are actually collected, we can significantly improve the quality and availability of the data. Most scientists can probably think of a field campaign where the data are no longer available. NSIDC worked to avoid this problem by working closely with the investigators conducting the Cold Land Processes field experiment in the Colorado Rocky Mountains during the winter and spring of 2002 and 2003 (see [30]). Not only was NSIDC involved in the planning of the data collection, but also provided data technicians who worked closely with field surveyors during the experiment. These data technicians learned the data collection protocol with the surveyors, helped collect some of the data, and entered the data into computers the night after they were collected. By learning the protocol and immediately entering the data, technicians were able to identify missing values and anomalies in the data and run some automated quality control checks. They were then able to follow up with the surveyors soon after they collected the data to correct specific problems and to improve later data collection. Technicians were also able to provide the data to the lead scientists for immediate assessment. Overall, this led to a 10 to 20 percent improvement in data quality [31].

NSIDC has had similar experience with recent satellite remote-sensing missions. NSIDC is the archive for all the data from NASA’s Advanced Microwave Scanning Radiometer (AMSR) and Global Laser Altimetry System (GLAS). Although NSIDC was not directly involved in the acquisition of the data, it did work closely with the mission science and instrument teams well before the instruments were even launched. This allowed the data managers to have a much greater understanding of the engineering aspects of the data and the algorithms used to produce the higher-level products. The result is much better documentation and much earlier data availability. Data from both of these missions were available to the public only months after launch, in contrast to years with some historical systems where data managers were not involved until well after their launch (e.g, sea ice data from SSM/I).

There is nothing inherent about distributed data systems that should preclude early involvement of data managers, but again this is something to consider in the design of those systems. Furthermore, data manager involvement could be more difficult if traditional data management organizations are not directly involved in the distributed data system.

6 Conclusions

The scientific method requires that experimental results be reproducible. That means the data used in the original experiment must be available and understandable. Furthermore, reexamination of an early data set often can yield important new results.

Maintaining access to and understanding of scientific data sets has been a challenge throughout history. The trend to a more geographically distributed data management model may improve data access in the short run but raises additional challenges. We should be able to address many of these challenges by developing new tools and data management systems, but we must not forget the human component. Experience and a review of the known data management issues show that we achieve the greatest success in long term data stewardship only when there is a close collaboration between data providers, data users, and professional data stewards. As we move forward, we need to ensure that new technologies and new data archive models enhance this collaboration.

[1] NOAA's National Geophysical Data Center. “About the World Data Center System.” December 29, 2003. http://www.ngdc.noaa.gov/wdc/about.shtml. January 2004.

[2] National Research Council. 2003. “Government Data Centers: Meeting Increasing Demands.” The National Academies Press.

[3] Ad Age Group. “Data Center.” January 6, 2004. http://www.adage.com/datacenter.cms. Data Center. January 6, 2004.

[4] Texas Cancer Council. “Texas Cancer Data Center.” December 23, 2003. http://www.txcancer.org/. Texas Cancer Data Center. January 6, 2004.

[5] CCSDS. 2002. “Reference Model for an Open Archival Information System (OAIS).” CCDSD 650.0-B-1. Blue Book. Issue 1. January 2002. [Equivalent to ISO 14721:2002].

[6] Barnum, George D. and Steven Kerchoff. “The Federal Depository Library Program Electronic Collection: Preserving a Tradition of Access to United States Government Information.” December 2000. http://www.rlg.org/events/pres-2000/barnum.html. January 7, 2003.

[7] Moore, Reagan W. October 7, 1999. “Persistent Archives for Data Collections.” SDSC Technical Report sdsc-tr-1999-2.

[8] Ullman, Richard. “HDF-EOS Tools and Information Center.” 2003. http://hdfeos.gsfc.nasa.gov/hdfeos/index.cfm. January 7, 2004.

[9] Abrams, Stephen L. and David Seaman. "Towards a global digital format registry." World Library and Information Congress: 69th IFLA General Conference and Council, Berlin, August 1-9, 2003 http://www.ifla.org/IV/ifla69/papers/128e-Abrams_Seaman.pdf

[10] Shepard, T. and D. MacCarn. 1999. *The Universal Preservation Format: A Recommended Practice for Archiving Media and Electronic Records.* WGBH Educational Foundation. Boston. 1999.

[11] Hunolt, Greg. *Global Change Science Requirements for Long-Term Archiving. Report of the Workshop, Oct 28-30, 1998.* USGCRP Program Office. March 1999.

[12] SEEDS Formulation Team. Strategic Evolution of Earth Science Enterprise Data Systems (SEEDS) Formulation Team final recommendations report. 2003. http://lennier.gsfc.nasa.gov/seeds/FinRec.htm. Jan. 2004.

[13] The Cedars Project. “Cedars Guide to The Distributed Digital Archiving Prototype.” March 2002. http://www.leeds.ac.uk/cedars/guideto/cdap/. December 2003.

[14] “Invest to Save: Report and Recommendations of the NSF-DELOS Working Group on Digital Archiving and Preservation.” 2003. Prepared for the National Science Foundation’s (NSF) Digital Library Initiative and the European Union under the Fifth Framework Programme by the Network of Excellence for Digital Libraries (DELOS).

[15] “It’s About Time: Research Challenges in Digital Archiving and Long-Term Preservation, Final Report, Workshop on Research Challenges in Digital Archiving and Long-Term Preservation.” August 2003. Sponsored by the National Science Foundation, Digital Government Program and Digital Libraries Program, Directorate for Computing and Information Sciences and Engineering, and the Library of Congress, National Digital Information Infrastructure and Preservation Program, August 2003.

[16] Lachman, B.E., A Wong, D. Knopman, and K. Gavin. 2002. “Lessons for the Global Spatial Data Infrastructure: International case study analysis.” Santa Monica, CA: RAND.

[17] Rhind, D. 2000. Funding an NGDI. In Groot, R. and J. McLaughlin, eds. 2000. *Geospatial data and infrastructure: Concepts, cases, and good practice*. Oxford University Press.

[18] PDG (Panel on Distributed Geolibraries, Mapping Science Committee, National Research Council). 1999. *Distributed geolibraries: Spatial information resources*. Washington, DC: National Academy Press.

[19] Federal Geographic Data Committee. Revised June 1998. "Content Standard for Digital Geospatial Metadata." Washington, D.C.

[20] Clinton, W. 1994. *Coordinating geographic data acquisition and access to the National Spatial Data Infrastructure, Executive Order 12906*. Washington, DC. Federal Register 59 17671-4. 2pp.

[21] The CEDARS Project. 2001. "Reference Model for an Open Archival Information System (OAIS)." http://www.ccds.org/documents/pdf/CCSDS-650.0-R-2.pdf. December 2003.

[22] National Libarary of Australia. 1999. "Preservation Metadata for Digital Collections." http://www.nla.gov.au/preserve/pmeta.html". December 2003.

[23] NEDLIB. 2000. "Metadata for Long Term Preservation." http://www.kb.nl/coop/nedlib/results/preservationmetadata.pdf. December 2003.

[24] OCLC. 2001. "Preservation Metadata Element Set – Definitions and Examples." http://www.oclc.org/digitalpreservation/archiving/metadataset.pdf. December 2003.

[25] The OCLC/RLG Working Group on Preservation Metadata. 2002. "Preservation and the OAIS Information Model – A Metadata Framework to Support the Preservation of Digital Objects." OCLC Online Computer Library Center, Inc.

[26] ISO Technical Committee ISO/TC 211, Geographic Information/Geomatics. "May, 2003. "Geographic information – Metadata. ISO 19115:2003(E)." International Standards Organization.

[27] ISO Technical Committee ISO/TC 200. November 2002. "Scope." http://www.isotc211.org/scope.htm#19139, January 2003.

[28] Holdsworth, D. "The Medium is NOT the Message or Indefinitely Long-Term Storage at Leeds University." 1996. http://esdis-it.gsfc.nasa.gov/MSST/conf1996/A6_07Holdsworth.html. January 2004.

[29] Stroeve, J, X. Li, and J. Maslanik. 1997. "An Intercomparison of DMSP F11- and F13-derived Sea Ice Products." NSIDC special report 5. http://nsidc.org/pubs/special/5/index.html. January 2004.

[30] Cline, D. et al. 2003. Overview of the NASA cold land processes field experiment (CLPX-2002*). Microwave Remote Sensing of the Atmosphere and Environment III*. Proceedings of SPIE. Vol. 4894.

[31] Parsons, M. A., M. J. Brodzik, T. Haran, N. Rutter. 2003. *Data management for the Cold Land Processes Experiment*. Oral presentation, 11 December 2003 at the meeting of the American Geophysical Union.

[32] The CEDARS Project. “CEDARS Guide to: Digital Preservation Strategies.” April 2, 2002. http://www.leeds.ac.uk/cedars/guideto/dpstrategies/dpstrategies.html. January 2004.

[33] National Library of Australia. “Persistent identifiers – Persistent identifier Scheme Adopted by the National Library of Australia.” September 2001. http://www.nla.gov.au/initiatives/nlapi.html. January 2004.

[34] RLG/OCLC Working Group on Digital Archive Attributes. May 2002. “Trusted Digital Repositories: Attributes and Responsibilities.” RLG Inc.

[35] Brodie, N. December 2000. “Authenticity, Preservation and Access in Digital Collections.” Preservation 2000: An International Conference on the Preservation and Long Term Accessibility of Digital Materials. RLG Inc.

[36] Ashley, K. December 2000. “I’m me and you’re you but is that that?” Preservation 2000: An International Conference on the Preservation and Long Term Accessibility of Digital Materials. RLG Inc.

[37] Lynch, C. 2000. “Authenticity and Integrity in the Digital Environment: An Exploratory Analysis of the Central Role of Trust.” Council on Library and Informational Resources, Washington D.C.

[38] Thibodeau, K. 2002. “Overview of Technological Approaches to Digital Preservation and Challenges in Coming Years.” The State of Digital Preservation: An International Perspective. Conference Papers and Documentary Abstracts. http://www.clir.org/pubs/reports/pub107/thibodeau.html. December 2003.

[39] Lee, K., Slattery, O., Lu, R., Tang X., McCrary V. 2002. *The State of the Art and Practice in Digital Preservation*. J. Res. Natl. Inst. Stand. Technol. 107, 93-106.

[40] DoD 8320.1-M. "Data Administration Procedures." March 29, 1994. Authorized by DoD Directive 8320.1. September 26, 1991. *Reference*: DoD 8320.1-M-1, "Data Element Standardization Procedures," January 15, 1993.

[41] International Permafrost Association, Data and Infromation Working Group, compilors. 1998. *Circumpolar active-layer permafrost system, version 1.0*. Boulder, CO: National Snow and Ice Data Center/World Data Center for Glaciology. CD-ROM.

[42] International Permafrost Association Standing Committee on Data Information and Communication, compilors. 2003. *Circumpolar active-layer permafrost system,*

version 2.0. Edited by M. Parsons and T. Zhang. Boulder, CO: National Snow and Ice Data Center/World Data Center for Glaciology. CD-ROM.

[43] Diomidis Spinellis. The decay and failures of web references. 2003. Communications of the ACM, 46(1):71-77.

[44] ICTU. "Digital Preservation Testbed." Digitale Duurzaamheid. 2004. http://www.digitaleduurzaamheid.nl/home.cfm Jan. 2004.

[45] Diamond, H., Bates, J., Clark D., Mairs R. "Archive Management – The Missing Component." April 2003, http://storageconference.org/2003/presentations/B06_Jensen..pdf, NOAA/NESDIS. January 2004.

[46] Hedstrom, M. 2001. "Digital Preservation: A Time Bomb for Libraries." http://www.uky.edu/~kiernan/DL/hedstrom.html. Jan. 2004.